BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Knowledge-based prediction of protein structures and the design of novel molecules

T. L. Blundell, B. L. Sibanda, M. J. E. Sternberg & J. M. Thornton

Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Prediction of the tertiary structures of proteins may be carried out using a knowledge-based approach. This depends on identification of analogies in secondary structures, motifs, domains or ligand interactions between a protein to be modelled and those of known three-dimensional structures. Such techniques are of value in prediction of receptor structures to aid the design of drugs, herbicides or pesticides, antigens in vaccine design, and novel molecules in protein engineering.

TECHNOLOGICAL developments of industrial, clinical and agricultural importance may be achieved in the coming years by imitation of the interactions between macromolecules and ligands that occur naturally in the living cell. For example, the design of drugs, herbicides and pesticides may be improved from knowledge of the interaction of a molecule with an isolated receptor, enzyme or nucleic acid. The specificity, stability or activity of engineered hormones of clinical importance or enzymes of value to the chemical industry may be improved from knowledge of proteins in general[1]. New peptide and protein vaccines may also be designed from information on antibody–antigen binding[2]. In the longer term new molecular electronic devices may be constructed using novel molecules—perhaps proteins—that aggregate in a predetermined way (as they do in living cells) and provide a template for molecules that conduct electrons; these will be the new biological microchips[3].

All these processes require information on the shape, charge, chemical function and dynamical flexibility of at least two interacting molecules. In most cases, at least one of the molecules will be a protein—enzyme, receptor, immunoglobulin, redox protein or polypeptide hormone. Much depends on knowledge of detailed three-dimensional structures defined by X-ray analysis, although medium resolution structures of small proteins in non-crystalline environments will increasingly come from two-dimensional NMR techniques[4] in the future. However, our knowledge of three-dimensional structures of proteins defined by these methods has increased only slowly, in spite of the greater number of laboratories involved in this work.

In contrast, the advent of recombinant DNA techniques has led to an explosion of information concerning the sequences of receptors and enzymes important for drug, herbicide and pesticide design. At the same time site-specific and deletion mutagenesis offer new possibilities of engineering new proteins with point mutations, with hybrid domains and even with new functions[5]. Very few proteins that have been sequenced and can now be expressed and manipulated in the laboratory have known three-dimensional structures, and so rational approaches to design are difficult. However, computer technology has also developed at a rapid rate and offers some compensatory possibilities. These include the use of database technology to store, retrieve and compare the known sequences; computer graphics to display models and manipulate known three-dimensional structures[6]; and computer simulation to calculate low-energy conformers, normal modes and molecular dynamics[7]. Together these new technologies offer the chance of exploiting our experimental knowledge of the three-dimensional structures of proteins in a rational approach to the modelling of protein interactions and the design of novel molecules.

In this review we emphasize our own view that modelling of proteins and their interactions is most usefully carried out using a knowledge-based approach. This depends on the identification of analogies between a protein that is to be modelled and other proteins of known three-dimensional structure at all levels in the hierarchy of protein organization: secondary structure, motifs, domains and quaternary or ligand interactions.

Sequence alignment

The first step in any modelling is to convert the DNA sequence available into the primary structure of the gene product and search for sequence homologies with known sequences of other proteins. Information on DNA sequences is available in databanks such as those of the European Molecular Biology Laboratory (EMBL), Heidelberg, of the National Biomedical Research Foundation at Maryland, United States, or GenBank (Bolt, Beranach and Newman Inc.) (4,000,000 bases and 5,000 entries). Protein sequences are collected together in the Protein Information Resource (PIR) databank at National Biomedical Research Foundation, Maryland, or the NEWAT databank in the USA. The PIR databank now contains over 3,000 protein sequence entries, many of which have been derived by translation of DNA sequences (for a review of sequence databases, see ref. 8). The National Biomedical Research Foundation provides some software to search their databanks, to compare user-specific segments with segments of the same length in the database, to align two sequences, to detect similarities and to score and display the degree of similarity.

The sequence alignment algorithms of Needleman and Wunsch[9], Sellers[10] and Waterman *et al.*[11] use dynamic programming methods to carry out a global comparison of two sequences. Their success depends very much on the degree of similarity and may give variable results depending on the gap-penalty parameters chosen, even for closely related sequences[12,13]. For sequences that are similar (>25%) such automatic procedures will identify the homology above the background of randomized sequences. As an approximate guide, if the alignment score is more than six standard deviations above that for random alignment, then most residues in secondary structures will be correctly aligned[13]. Because insertions and deletions often occur at the loop regions of proteins between secondary structures, improvements can be made by introducing penalties for insertions/deletions in α-helices or β-strands[13,14].

For very distantly related sequences, homology may be restricted to a few key residues or sequence segments whose separation along the chain may vary considerably between proteins. There have been several developments of algorithms to identify such local homologies[15–18].

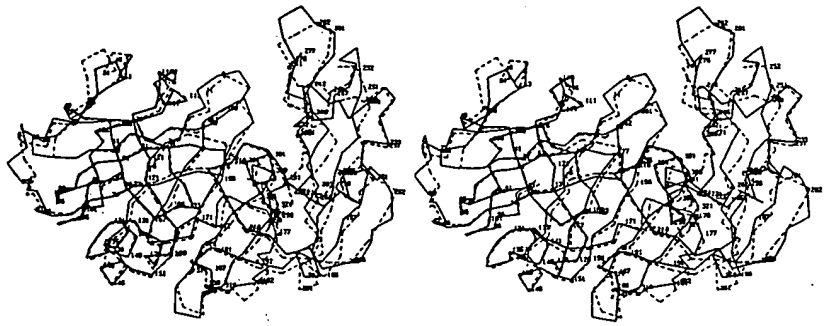


Fig. 1 A stereo view comparing the three-dimensional structures of two aspartic proteinases, endothiapepsin and penicillopepsin, indicates that the central cores, and active-site-cleft residues are closely similar. Diversity is achieved mainly in the loop regions that occur at the periphery of the bilobal enzymes.

The tertiary structures of homologous proteins

Comparison of the tertiary structures of homologous proteins shows that three-dimensional structures are conserved in evolution more than protein primary structures and considerably more than DNA sequences (ref. 19 for review). In recently diverged molecules the elements of secondary structure—α-helices and β-strands—are arranged in closely comparable three-dimensional topologies. Amino-acid replacements occur most often in surface positions so that the main chain conformations are little affected[20]. More radical insertions and deletions tend to occur in surface loops between the secondary structure units although insertions are allowed in some β-strands to give rise to a β-bulge[21]. In some families of proteins such as the insulin family this divergence has occurred with complete conservation of the hydrophobic core[22].

However, in most protein families divergence is accompanied by changes in the hydrophobic core. This has been examined in an elegant series of papers by Chothia and Lesk[23-25] who have shown that although the volume of the core remains approximately constant in many families, amino-acid replacements of hydrophobic core residues are usually accommodated not by complementary amino-acid substitutions but rather by small shifts of secondary structural elements. They have shown that in the α-helical globins large relative shifts of certain helices are observed, but in the β-sheet family of azurins/plastocyanins relative rotations of the β-strands have occurred. To a certain extent this appears to be family dependent and is less evident not only in the immunoglobulins where an intersheet disulphide bridge pins them together but also in the core of other β-sheet proteins such as aspartic proteinases (Fig. 1).

What is the extent of such differences in the hydrophobic cores? To define this quantitatively, the topological equivalence of residues in homologous proteins is identified by a least-squares procedure that fits the proteins in three dimensions and calculates the root-mean-square (r.m.s.) differences of equivalent C_α atoms. A problem arises in the definition of which residues constitute the core. Chothia and Lesk[26] first superposed the major elements of secondary structure and then extended them to include additional residues at each end so long as the deviations did not exceed 3 Å. Using this definition the 'common core' included over 90% of the amino acids for proteins of >50% identity. At 50% identity the r.m.s. deviation averaged ~1 Å but there was a correlation between r.m.s. deviation and percentage residue identity of the complete protein.

This definition of the common core may not always be most appropriate. If the core is defined by those residues whose side chains are inaccessible to solvent, a similar correlation is found[27] although the number of residues and the average r.m.s. deviations are much smaller (Fig. 2).

Some of the differences between homologous structures arise from crystallographic errors and from conformational differences in the different crystalline environments. This can be estimated from comparison of proteins of identical sequence whose tertiary structures have been defined independently. The hydrophobic cores may differ by only 0.2 Å in well defined, high resolution (<1.8 Å) structures but for all residues the differences are likely to be 0.3–0.6 Å. Most significantly the differences can be 1 Å for structures at intermediate (2.5–3.0 Å) resolutions.

Modelling by homology

Given homologous proteins with three-dimensional structures defined by high resolution X-ray analysis, how should we proceed to construct a model? Historically, this has been achieved in several simple stages starting with alignment of the sequences, followed by creation of insertions, deletions and replacements in the known three-dimensional structure of the homologous protein. This was first carried out in 1969 by Browne and co-workers[28] in the construction of a model of α-lactalbumin based on the three-dimensional structure of lysozyme which had been defined by X-ray analysis. Although it was achieved using physical models, subsequent attempts exploited computer graphics as the technology developed, especially the program FRODO[29]. The initial models have been refined using energy minimization techniques to give final structures without steric clashes[30]. Molecules modelled in this way have included

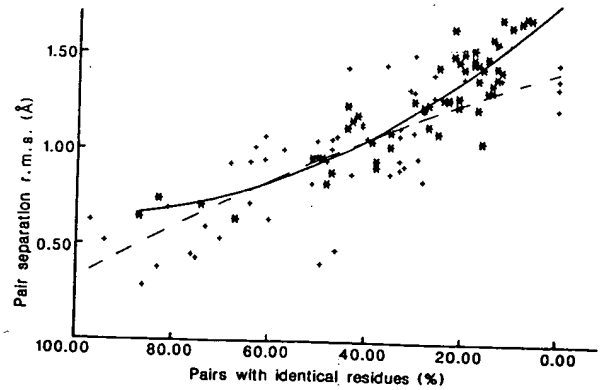


Fig. 2 A graph in which stars (*) indicate the root-mean-square (r.m.s.) separation of topologically equivalent residue (Cα) positions in pairs of proteins plotted as a function of the degree of identity of equivalent residues that are <3 Å in separation when the two proteins are arranged as a best fit. The best fitting curve is given as a continuous line. Crosses (+) show residues whose side chains are <7% accessible to the solvent and the best-fitting curve is given by a dashed line. The percentage identities refer only to the residues compared. For any particular protein, the percentage homology is increased but the r.m.s. separation and the number of residues are decreased when the solvent-inaccessible residues are compared. However, apart from proteins of low homology, the relationship between the r.m.s. separation and sequence homology is similar in each analysis (T. Hubbard, personal communication).

relaxins[31,32], insulin-like growth factors[33], serine proteinases[20], HLA-DR antigens[34], angiogenin[30], aspartic proteinases such as renin[35–38] and immunoglobulins[39,40] (ref. 41 for review).

Modelling using multiple structures

With over 300 structures in the Brookhaven databank, of which only ~100 are non-homologous, there may well be more than one structure available to serve as a basis for modelling. One approach to this problem[42] is to model the structure on the basis of each available known tertiary structure and to test the resulting models for packing of side chains and their solvent accessibilities. Thus, calmodulin was constructed on the basis of intestinal calcium-binding protein and carp parvalbumin[42]. Although each sequence had approximately the same sequence homology with calmodulin, certain features of the model based on intestinal calcium-binding protein led to this being the preferred model.

An alternative knowledge-based approach is to use simultaneously but selectively the information from all the known tertiary structures of the homologous family. For example there are five high-resolution mammalian serine proteinase structures defined by X-ray analyses. The first step is to align the sequences and tertiary structures of the homologous proteins taking care that all structures are evenly weighted; it is not satisfactory to fit by least squares criteria all the tertiary structures to one of the family. One approach is to construct a 'framework' that contains a virtual atom at points in three-dimensional space of the topologically equivalent residues common to the family[43]. The second step is to align the new sequence with those of the family so that topologically equivalent residues of the 'framework' are equivalent in the sequence alignment. This generally involves some realignment from comparisons of the sequence alone; it may be achieved by hand or by using the template sequence routines of Taylor[18]. Fragments of each of the homologous proteins which are closest in sequence are selected; generally it is preferable to select fragments which join one secondary structure to another and if possible link one element of the framework to the other. For the modelling of the serine protease domain of tissue plasminogen activator on the basis of the then-available trypsin, elastase, chymotrypsin and kallikrein structures, seventeen fragments were used (Fig. 3). A similar procedure has been developed independently by Levitt and co-workers[39].

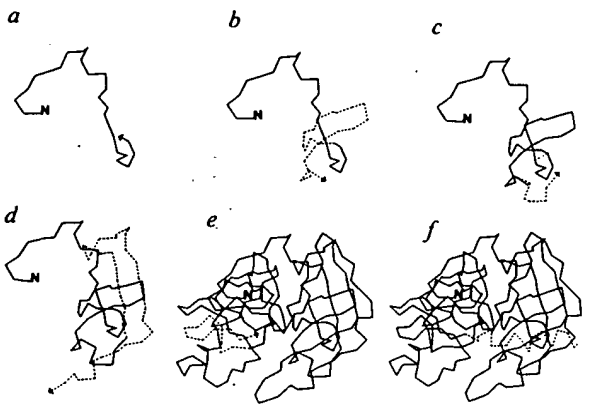


Fig. 3 A model of tissue plasminogen activator (*f*) constructed on the basis of four other mammalian serine proteases of defined three-dimensional structure. Seventeen fragments are selected on the basis of sequence homology. The first in the sequence is shown in *a* and *b*, *c* and *d* show the addition of subsequent fragments as dotted lines: *e* and *f* show the last two fragments in the sequence. Four further fragments are added from other proteins (D. Athwal, and T. Harris personal communication; T.L.B., unpublished data).

Insertions and deletions in loop regions

Because the majority of significant differences in protein structure occur in loop regions, these are the most difficult to construct. The choice of a conformation found in a loop of equivalent length in a homologous protein is often a good guide, even if the sequence differs[20], although this will occasionally be misleading[44,45]. Nevertheless, the problem remains how to identify loop sequences of the same length where conformations differ. If there is no sequence of equivalent length in any homologous protein, conformations from other proteins in similar supersecondary structures can be explored[1].

The database for such prediction of conformation in loops is now being constructed. Sibanda and Thornton[46] and Milner-White and Poet[49] have detailed the conformations of β-hairpin structures (loops between two adjacent antiparallel β-strands) and similar analyses are proceeding for loops between two α-helices and between α-helices and β-strands[47]. Detailed analysis of β-hairpin loops shows that in short loops certain structures with characteristic sequences recur many times in unrelated proteins (ref. 46; Fig. 4 and Table 1). In general there are few longer loops; the exceptions often include prolines and the sequence Pro-Gly is common.

Although there is little sequence homology between loops, certain glycines are often conserved and there is a preference at certain positions for residues that can help form β-strands or α-helices. For modelling we can then adopt a systematic approach that involves first identifying the loop length, secondly selecting a conformer on the basis of sequence and thirdly testing the conformer by least-squares fitting to the β-strands of the model framework. Similar techniques are applicable in principle for modelling all types of loops.

The power of this technique is shown in models of mouse and human renins[35,36] and chymosin[50] that were modelled on the basis of endothiapepsin. For example, sequences of equivalent loops in endothiapepsin and chymosin are given as:

Residue number	197	198	199	200	201	202	203	204
Endothiapepsin	Y	A	V	G	S	G	T	F
Chymosin	V	T	I	S	G	V	V	V

where the turns are underlined. In endothiapepsin this is a two residue type II′ turn with glycine at the first position. The sequence of chymosin with a glycine at the second position indicates that this is most probably a type I′ turn. In a second example a deletion of one residue is observed. Thus:

Residue number	238	238A	239	240	241	242	243	244	245	246	247
Endothiapepsin	A	K	S	S	S	S	V	G	G	Y	V
Chymosin	A	T	Q	N	Q	Y	—	G	E	F	D

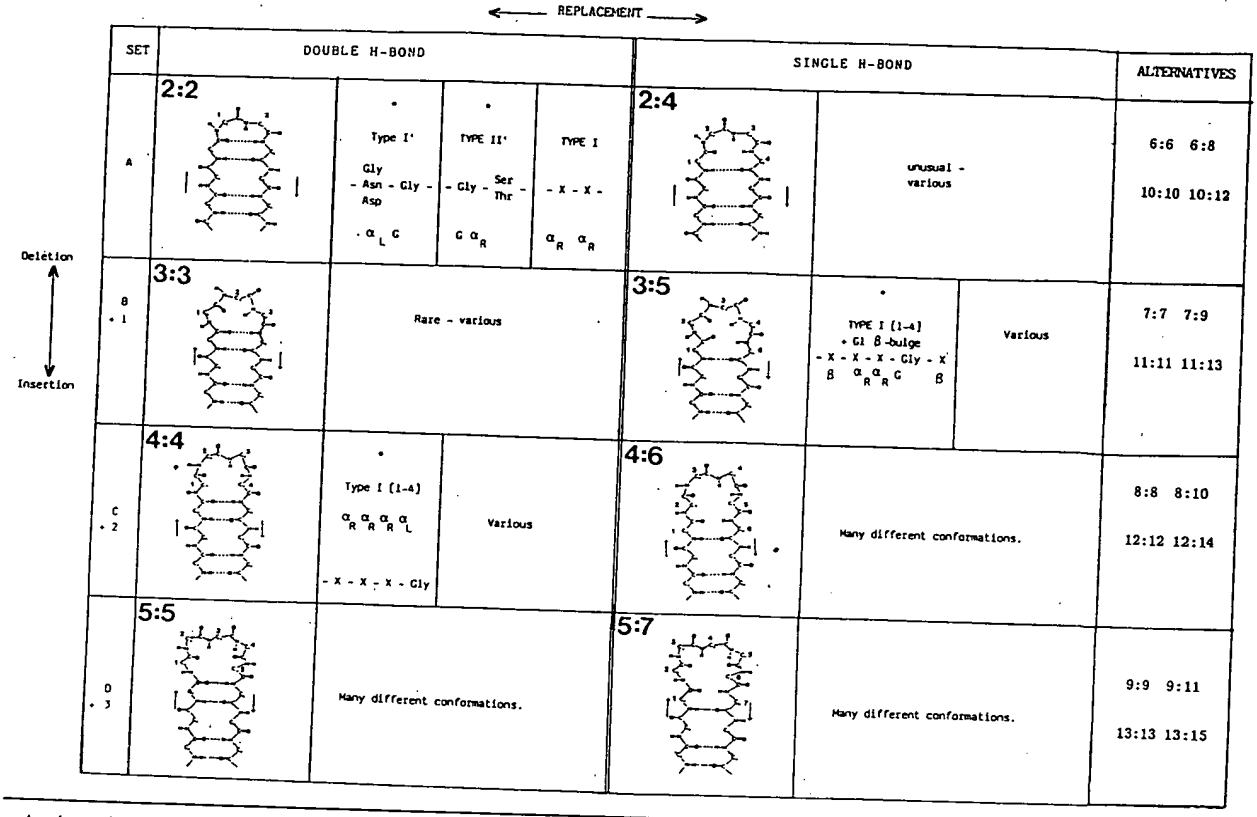
In endothiapepsin this is a standard four-residue loop with a glycine at position 4. In chymosin there is a deletion of one residue leading to a three-residue loop, but the presence of glycine implies that this may lose one hydrogen bond to give a standard five-residue loop with a glycine at position 4 (Fig. 5).

A similar approach has been adopted by Jones and Sirup[51] in their extension of FRODO. They search a structural database to find all possible loops of the correct length using an elegant method based on the distance between residues displayed as a matrix. A computer search for suitable motifs involves pattern matching of distance matrices. Alternatively the appropriate sequences and conformations can be identified in a relational database[1].

Where no structure appropriate for the modelling exists in conformations of proteins defined by high resolution X-ray analysis, we must adopt an alternative *ab initio* approach such

Table 1 Systematic modelling of β-hairpins

SYSTEMATIC MODELLING OF β-HAIRPINS



A schematic representation of possible β-hairpin structures found in proteins defined by X-ray analysis. A similar analysis has been carried out by Milner-White and Poet[49]. Structures in horizontal lines (such as 2:2 and 2:4) have identical numbers of residues and represent differing conformations that may be preferred by particular amino acid sequences; these are useful for modelling replacements. Structures in columns represent hairpin conformations with differing numbers of residues and these are used in modelling insertions or deletions. Loss of the top hydrogen bond implies a horizontal movement to the right (3:3→3:5) whereas loss or gain of two hydrogen bonds implies a change such as 3:5→7:9 which belongs to the same class. Preferred sequence patterns are indicated (from ref. 46 and B.L.S., J.M. and T.L.B. in preparation). Loop lengths are defined by two numbers X; Y where X is the number of residues not involved in β-strands and Y is the number of residues that do not have both amide and carboxyl groups involved in H-bonds characteristic of β-strands.

as use of a distance geometry algorithm[52] or a systematic search procedure of low energy conformers using molecular dynamics[48].

Having created the appropriate backbone, the amino acids that differ between known structures and the modelled structure must be replaced. For this we assume that the side-chain torsion angles for equivalent bonds are the same and that the side chain occupies a similar region in three-dimensional space in the family of structures. A set of rules for replacement of residues based on conformations of side chains at topologically equivalent positions in homologous proteins (M. Sutcliffe, personal communication) has been derived.

Energy minimization and molecular dynamics

Where the proteins have sequence homology of 50% or more, the models predicted by the methods described here will be probably correct to better than 1 Å although individual side chains may be more in error. Some but not all the errors can be removed by minimization of the potential energy using standard programs such as AMBER[53] or CHARMM[54,55]. Whereas X-ray analysis gives a time- and-space-averaged structure in a crystalline lattice with perhaps 40% solvent, energy minimization gives the structure at a single minimum usually *in vacuo* as solvent is not often simulated. Efforts to simulate the aqueous environment using Monte Carlo methods and perhaps to model specifically tightly bound waters[56] are particularly helpful for surface side chains although some contraction of the protein volumes tends to occur[57] unless the molecule and solvent are included in a repeating lattice. Use of energy minimization in this way finds only a local minimum and may be expected to be useful if the errors are relatively small (usually <1 Å).

To explore local changes requiring larger shifts, methods of constraining the major part of the structure whilst allowing freedom in the region of interest have been explored[55]. With the extent of the calculation reduced, side-chain torsion angles can be varied to locate conformations for subsequent energy minimization. Alternatively, energy calculations or molecular dynamics may often be useful to explore a range of local conformations, as developed by Robson and co-workers[58]. Energy minimization used in the conventional way will not be useful as a local energy minimum can always be found. No one has modelled successfully changes such as those identified by Lesk

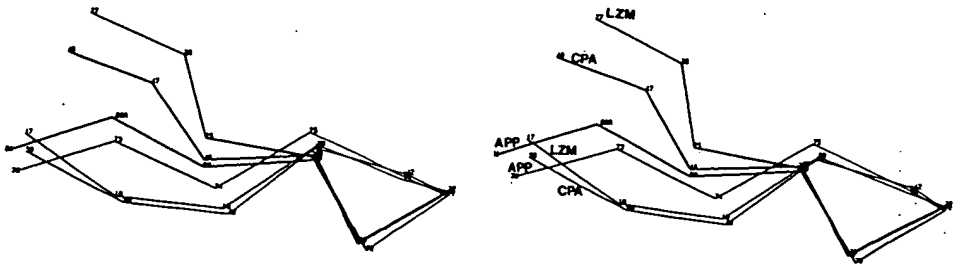


Fig. 4 A stereo view of three similar β-hairpin structures (of type 3:5, see Table 1) superposed with r.m.s. error (calculated for the five residues in the loop) of 0.56 Å. The examples given are from penicillopepsin (APP: residues 72–82), carboxypeptidase A (CPA: residues 38–48) and lysozyme T_4 (LZM: residues 17–27).

and Chothia[23] in the globins where an α helix can differ in position by 7 Å and rotational angle by 30°. The best chance appears to be in a simplified molecular dynamics procedure with rigid secondary structural units, or an interactive docking procedure such as that of Wodak *et al.*[57].

How correct are the models?

Several modelled proteins have been defined subsequently by X-ray analysis. The first of these models, that of α-lytic proteinase, was flawed by an incorrect sequence alignment. Others have been more successful. For example, the three-dimensional structures of rat γ3-1 crystallin and the orthologous protein γIV from calf were modelled on the basis of γII-crystallin[59]. The three-dimensional structure of the calf γIV defined by X-ray analysis at 2.3 Å resolution has been shown to have a r.m.s. difference of ~0.7 Å with the model overall and 0.5 Å in each domain[60]. Chothia *et al.*[39] have modelled the variable domain of immunoglobulin and have correlated their prediction with the recently defined X-ray analysis of Poljak and co-workers[61]. They have not only predicted the loop conformations correctly in 4 of the 6 hypervariable regions, but also attained a r.m.s. difference of 1 Å in the framework region for the main chain.

When a structure is modelled on the basis of one homologous protein, the relationship of the percentage sequence identity to the r.m.s. difference given by Chothia and Lesk is a reasonable guide to the precision. A model of sperm whale myoglobin based on either human or equine β-chains of haemoglobin has an r.m.s. difference of 1.40 Å from the structure defined by X-ray analysis. The precision can be improved by fitting the modelled structure to the framework derived from all the homologous structures. This leads to a smaller r.m.s. difference of 1.25 Å between the model and the X-ray structure[43] when the α- and β-chains of equine and human haemoglobins are used.

Although the absolute value of potential energy of the model as calculated by programs such as CHARMM is not a good guide to its correctness[62], the distribution of the hydrophobic side chains and the nature of the solvent-accessible surface are more sensitive indicators; these are features that can be quantified but are also easily assessed visually. For example, a buried charged side chain that is not hydrogen-bonded is almost certainly an indication that the model is incorrect.

Perhaps the most useful indication of the correctness of a model is its ability to predict a molecular interaction that can be experimentally tested. Models of insulin-like growth factors[33] demonstrated that these molecules might bind insulin receptors. Electrostatic potential surfaces generated for the model of calmodulin have suggested a binding site consistent with experimental data for basic amphiphilic α-helical peptides[42]. The size and chemical nature of the predicted combining sites of anti-lysozyme monoclonal antibodies have been tested and are consistent with epitope boundaries[40]. The size, shape and position of the specificity pockets of human renin[36–38] have been assessed as useful predictors for the design of inhibitors.

Future developments

There are two challenges for the future. First, to extend the method to cases where no obvious sequence homology may be found, but where the unknown structure is probably constructed

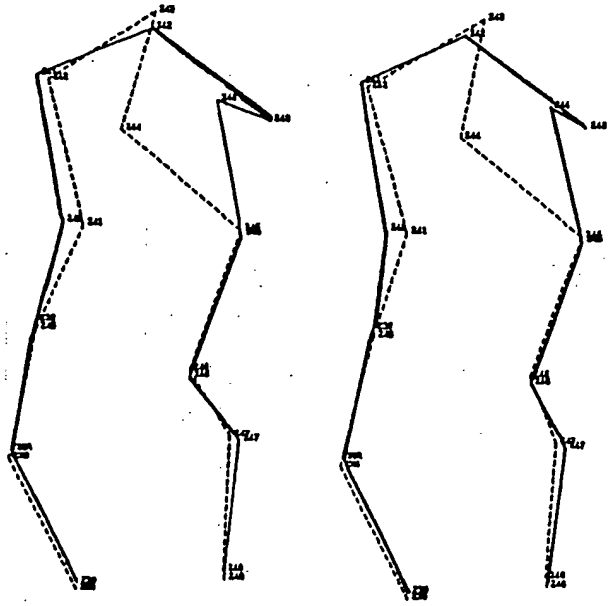


Fig. 5 A model of chymosin residues 238–248 which form a β-hairpin (broken line). The model is based on the high-resolution refined structure of endothiapepsin (continuous line). A single residue deletion in the chymosin structure results in a remodelling of the family 4:4 β-hairpin of endothiapepsin as a family 3:5 hairpin in chymosin.

from the structural motifs, $\beta\alpha\beta$ units, $\beta\beta$ hairpins, an 8-fold $\alpha\beta$ barrel such as found in triose phosphate isomerase[63] or a jelly-roll structure, such as is found in the viruses[64]. The problem is to identify these motifs from the sequence through detailed study of the critical sequence patterns for each motif and to establish the sequence fingerprint against which a sequence of unknown structure can be compared.

For example, a particular pattern of a glycine and a serine along with a few conservatively varied residues and hydrophobic regions in a sequence of ~40 residues has been used to identify a Greek-key structure in the surface protein S of the sporulating bacterium *Myxococcus xanthus* even though there is no significant overall homology[65]. Another pattern Gly-X-Gly-X-X-Gly, where X is any amino acid, has been useful for identification of an α/β structure in the p21 protein[66] and in tyrosyl kinases[67].

Second, we must use the methods to design novel molecules. For protein engineering replacements, small insertions and deletions, the approach is a natural extension of the modelling procedures. Sequences in similar structural motifs can be identified from known protein structures and grafted onto the known structure. Local conformations can be explored by global searches, restrained minimization approaches or molecular dynamics followed by energy minimization. In the design of hybrid molecules for example linking monoclonal antibody Fab fragments to enzymes such as proteases for dissolving blood-clots, or to toxins for killing tumour cells, the identification of stable domains and the design of linker regions will be of central importance[68]. In the design of novel tertiary structures which might be endowed with special binding or catalytic functions, the complete technology evolved for modelling will be essential.

In drug design the modelling of the receptor must be complemented with the definition of the ligand-binding interaction. Much is to be learnt from interactions in homologous systems such as other members of the enzyme family. Thus, a starting point for modelling human angiotensinogen interactions with renin is usefully obtained from high resolution X-ray studies of inhibitors with other aspartic proteinases such as endothiapepsin[1,69]. Also, detailed analyses of side chain–side chain and side chain–ligand interactions can define preferred orientations, for example of phenylalanine rings with oxygens[70], sulphur[71] and other aromatic groups[72].

For vaccine design, the synthetic molecule must mimic the antigen to stimulate the appropriate immune response[73,74]. Optimal activity will be achieved by designing the synthetic molecule to mimic the critical features of the native antigen as closely as possible. As in drug design it may be advantageous to restrict flexibility, for example by ring closure. Alternatively, an antigenic loop peptide may be grafted from one protein into another using the technique of protein engineering. Indeed it may be possible to produce a protein vaccine that incorporates antigenic peptides from many proteins and confers multi-valent immunity. For rational design of such complex vaccines and even simple epitopes, model building using powerful computer graphics including superposing, annealing and energy minimizing structures, will provide starting points to guide the experimenter.

We thank Willie Taylor, Ian Tickle, Suhail Islam, Michael Sutcliffe, Tim Hubbard, David Barlow, Dee Atwal, Ilyas Haneef, Andrew Hemmings, James Milner-White, Geoff Barton, Mark Edwards, Marketa Zvelebil, Jus Singh and Stephen Bryant for discussion of the ideas discussed in this review. We thank the SERC, ICI, CellTech, Glaxo and Sturge for financial support.

1. Blundell, T. L. *et al. Phil. Trans. R. Soc.* A 317, 333–344 (1986).
2. Porter, R. & Whelan, J. (eds) *Synthetic Peptides as Antigens, CIBA Foundation Symposium* 119 (Wiley, Avon, 1986).
3. Haddon, R. C. & Lamola, A. A. *Proc. natn. Acad. Sci. U.S.A.* 82, 1874–1878 (1985).
4. Wagner, G. & Wuthrich, K. *J. molec. Biol.* 155, 347–366 (1982).
5. Winter, G. & Fersht, A. R. *Trends Biotechnol.* 2, 115–119 (1984).
6. Richards, W. G. (ed.) *J. molec. Graphics* 4, 1–73 (1986).
7. Brooks, B. & Karplus, M. *Proc. natn. Acad. Sci. U.S.A.* 80, 6571–6575 (1983).
8. Kneale, G. G. & Bishop, M. J. *Cabios Rev.* 1, 11–17 (1985).
9. Needleman, S. B. & Wunsch, C. D. *J. molec. Biol.* 48, 443–453 (1970).
10. Sellers, P. *J. appl. Math.* 26, 787–793 (1974).
11. Waterman, M. S., Smith, T. F. & Beyer, W. A. *Adv. Math.* 20, 367–387 (1976).
12. Fitch, W. M. & Smith, T. F. *Proc. natn. Acad. Sci. U.S.A.* 80, 1382–2386 (1983).
13. Barton, G. & Sternberg, M. J. E. *Protein Engng* (in the press).
14. Lesk, A., Levitt, M. & Chothia, C. *Protein Engng* 1, 77–78 (1987).
15. Gooad, W. B. & Kanehisa, M. I. *Nucleic Acids Res.* 10, 247–263 (1982).
16. Sellers, P. *Proc. natn. Acad. Sci. U.S.A.* 76, 3041–3044 (1979).
17. Boswell, D. R. & McLachlan, A. D. *Nucleic Acids Res.* 12, 457–464 (1984).
18. Taylor, W. R. *J. molec. Biol.* 188, 233–258 (1986).
19. Bajaj, M. & Blundell, T. L. *A. Rev. Biophys. Bioengng* 13, 453–492 (1984).
20. Greer, J. *J. molec. Biol.* 153, 1027–1042 (1981).
21. Richardson, J. S., Getzoff, E. D. & Richardson, D. C. *Proc. natn. Acad. Sci. U.S.A.* 75, 2574–2578 (1978).
22. Cutfield, J. F. *et al.* in *Structural Studies on Molecules of Biological Interest* (ed. Dodson, G. C.) 501–508 (Clarendon Press, Oxford, 1981).
23. Lesk, A. M. & Chothia, C. *J. molec. Biol.* 136, 225–270 (1980).
24. Lesk, A. M. & Chothia, C. *J. molec. Biol.* 160, 325–342 (1982).
25. Chothia, C. & Lesk, A. M. *J. molec. Biol.* 160, 309–323 (1982).
26. Chothia, C. & Lesk, A. M. *EMBO J.* 5, 823–826 (1986).
27. Hubbard, T. & Blundell, T. L. *Prot. Engng* (submitted).
28. Browne, W. J. *et al. J. molec. Biol.* 120, 97–120 (1969).
29. Jones, A. T. *J. appl. Cryst.* 11, 268–272 (1978).
30. Palmer, K. A., Scheraga, H. A., Riordan, J. F. & Vallee, R. Z. *Proc. natn. Acad. Sci. U.S.A.* 83, 1965–1969 (1986).
31. Isaacs, N. *et al. Nature* 271, 278–281 (1978).
32. Bedarkar, S. *et al. Nature* 270, 449–451 (1977).
33. Blundell, T. L., Bedarkar, S. & Humbel, R. E. *Proc. natn. Acad. Sci. U.S.A.* 75, 180–184 (1978).
34. Travers, P. *et al. Nature* 310, 235–278 (1984).
35. Blundell, T. L., Sibanda, B. L. & Pearl, L. H. *Nature* 304, 273–275 (1983).
36. Sibanda, B. L. *et al. FEBS Lett.* 174, 103–111 (1984).
37. Carlson, W., Karplus, M. & Haber, E. *Hypertension* 7, 13–26 (1985).
38. Akahone, P. *et al. Hypertension* 7, 3–12 (1985).
39. Chothia, C. *et al.* Science 233, 755–758 (1986).
40. La Paz, P., Sutton, B. R., Darsley, M. J. & Rees, A. R. *EMBO J.* 5, 415–425 (1986).
41. Ripka, W. C. *Nature* 321, 93–94 (1986).
42. O'Neil, K. & De Grado, W. F. *Proc. natn. Acad. Sci. U.S.A.* 82, 4954–4958 (1985).
43. Sutcliffe, M. & Haneef, I. *Inf. Q. Prot. Cryst.* 18, 11–18 (1986).
44. Read, R. J., Fujinaga, M., Sielecki, A. R. & James, M. N. G. *Biochemistry* 22, 4420–4433 (1983).
45. Read, R. J., Brayer, G. D., Jarasek, L. & James, M. N. G. *Biochemistry* 23, 6570–6575 (1984).
46. Sibanda, B. L. & Thornton, J. M. *Nature* 316, 170–174 (1985).
47. Edwards, M., Sternberg, M. J. E. & Thornton, J. M. *Prot. Engng* (in the press).
48. Moult, J. & James, M. N. *Proteins* 2, 146–163 (1986).
49. Milner-White, J. & Poet, R. *Biochem. J.* 240, 289–292 (1986).
50. Sibanda, B. L. thesis, Univ. London (1986).
51. Jones, T. A. & Sirup, T. *EMBO J.* 5, 819–822 (1986).
52. Crippen, G. M. in *Chemometric Research Studies Series* Vol. 1 (ed. Bandan, D.) 1–58 (Wiley, New York, 1981).
53. Weiner, S. J. *et al. J. Am. chem. Soc.* 106, 765–784 (1984).
54. Brooks, B. R. *J. Computation Chem.* 4, 187–217 (1983).
55. Shih, H. L., Brady, J. & Karplus, M. *Proc. natn. Acad. Sci. U.S.A.* 82, 1697–1700 (1985).
56. Hagler, A. T. & Moult, J. *Nature* 272, 222–227 (1979).
57. Wodak, S. J., Alard, P., Delhouse, P. & Renneboog-Squilbin, C., *J. molec. Biol.* 181, 317–322 (1984).
58. Robson, B. & Platt, E. *J. molec. Biol.* 188, 259–281.
59. Summers, L. D. *et al. Expl Eye Res.* 43, 77–92 (1986).
60. Driessen, H. P. C. & White, H. in *Molecular Replacement* (ed. Machin, P. A.) 27–32 (Daresbury Laboratory, Daresbury, 1985).
61. Amit, A. G. *et al. Science* 233, 747–753 (1986).
62. Novotny, J., Bruccoleri, R. & Karplus, M. *J. molec. Biol.* 177, 787–818 (1984).
63. Banner *et al. Nature* 255, 609–614 (1975).
64. Abad-Zapatero, C. *et al. Nature* 286, 33–39 (1980).
65. Wistow, G., Summers, L. J. & Blundell, T. L. *Nature* 315, 771–773 (1985).
66. Wieranga, R. K. & Hol, W. G. *Nature* 302, 842–844 (1983).
67. Sternberg, M. J. E. & Taylor, W. R. *FEBS Lett.* 175, 387–392 (1984).
68. Neuberger, M. S., Williams, G. T. & Fox, R. O. *Nature* 312, 604–608 (1984).
69. Hemmings, A., Foundling, S., Sibanda, B. L., Wood, S. P. & Blundell, T. L. *Trans. Biochem. Soc.* 13, 1036–1041 (1985).
70. Thomas, K. A., Smith, G. M., Thomas, T. J. & Feldmann, R. J., *Proc. natn. Acad. Sci.* 79, 4843–4847 (1982).
71. Reid, K. S. C., Lindley, P. F. & Thornton, J. M. *FEBS Lett.* 190, 209–213 (1985).
72. Singh, J. & Thornton, J. M. *FEBS Lett.* 191, 1–6 (1985).
73. Nussenzweig, V. & Nussenzweig, R. in *Ciba Foundation Symposium* 119 (eds Porter, R. & Whelan, J.) 150–163 (Wiley, Avon, 1986).
74. Geysen, H. M., Rodda, S. J. & Mason, T. J. in *Ciba Foundation Symposium* 119 (eds Porter, R. & Whelan, J.) 130–149 (Wiley, Avon, 1986).